ASSIGNMENT 2.1  
**Sheryl Oilar**

**Due 9/30/13**

1. What are the advantages of using a SQL-based database with a cartridge for storing 2D chemical structures, over a system developed just for handling chemical structures (such as ISIS)?
2. Although the ISIS and other similar systems may be fast when searching for substructures, they only store information on the structure of the chemical and do not give any information on bioactivity or physical properties. An SQL relational database has many advantages because it can expand beyond structural features to include other information that can be used to query compounds such as physical properties or assay scores while retaining the ability to search by substructure through the use of a fingerprint. An SQL database is also much more flexible and easily modified without the need to change the structure of the entire database.
3. Assume we have a PostgreSQL table called **compounds** with the following fields:

SMILES VARCHAR(200)

Name VARCHAR(200)

FP BIT(166)

Activity REAL

* 1. Write a SQL statement that will populate the FP field with 166-bit MACCS keys using the gNova cartridge
     1. update compounds set FP=public166keys(SMILES);
  2. Write a SQL statement that will search for all records with an activity greater than 70 that contain a Pyridine ring (An aromatic ring with a Nitrogen in)
     1. SELECT \* FROM compounds

WHERE matches(smiles, ‘\*c1\*n\*c1\*’)

AND Activity > 70;

1. Explain in plain english what the following SMARTS means:  
   [R0;!#8]
2. This is an atom that is in zero rings AND is not Oxygen.
3. Can you write the SMARTS definitions for Lipinski's Hydrogen bond donor and acceptor? *Donors are defined as nitrogen or oxygen atoms that have at least one directly bonded hydrogen atom and Acceptors are defined as nitrogen or oxygen*
   * 1. Donor: [O,N;!H0]

Acceptor: [O,N]

1. What is the difference between Reaction SMILES and SMIRKS?
2. **SMIRKS** – Extension of SMARTS that depicts a generalized reaction mechanisms of the side chains or functional groups only before and after transformation in a chemical reaction.

**Reaction SMILES** – Extension of regular SMILES that enables simple representation of reactions with both reagents and products represented in whole chemical SMILESformat. Reaction SMILES enable mapping of atoms from reagents to products.

**Both** mechanisms are generalized and do not include details such as stoichometry, catalysts, conditions, or yield etc

ASSIGNMENT 2.2

1. Use any of the pubchem bioassay confirmatory screening data (Malaria/ Tuberculosis/ Leishmaniasis)
2. Download the active and inactive compounds using the pubchem API or the using [rpubchem](https://github.com/rajarshi/cdkr/tree/master/rpubchem/R) package in R.
3. If the dataset you select is small then its fine or else you can sample a 100 active compounds and 100 inactive compounds from a big dataset and use it for further analysis.
4. Use different fingerprints from [rcdk](http://cran.r-project.org/web/packages/rcdk/rcdk.pdf) to plot a heatmap of the tanimoto similarity matrix.
5. Describe briefly what can you infer from the data and post it on [cheminfoclub wiki](http://cheminfoclub.wikispaces.com/home)

Discussion: Using AID 588680 as my example data set I followed the posted video. The assay heatmap shows that most of the compounds used in the assay were fairly similar, with a Tanimoto value around 0.7. This leads me to believe the compounds were likely selected based on a targeted similarity for compounds with previous activity.